The new era of disinformation wars

Lisa M. Cohen 2020-11-30T00:00:00

While the manipulation of photographs has traditionally been deemed a State intelligence privilege, today's technological evolution allows anyone to effortlessly modify digital material – deepfakes being the newest, and arguably most dangerous, trend of such practices. Deepfake algorithms use so-called 'deep learning' artificial intelligence (AI) to create new audio and video by replacing or merging one's voice and/or face with manipulated and artificial data, which automatically fits the output dimensions and conditions. A short recording of one's voice suffices for an AI to create a "voice skin" that can be processed to say virtually anything. Although deepfakes are currently mainly used for humoristic purposes (see examples here), their use for malicious and military purposes seems inevitable. Indeed, deepfakes offer the potential to deceive and misinform adversaries and gain significant military advantages, while debunking and attributing the misinformation remains highly difficult. Against this background, and with social media spreading information to a massive community within seconds, feigning an alternative reality may set off an uncontrollable chain of events with detrimental consequences for the civilian population in conflict-ridden areas.

Aiming to determine whether international humanitarian law (IHL) sufficiently regulates the use of deepfakes, this Bofax examines if and how existing IHL norms apply to deepfakes, particularly against the backdrop of the 2017 Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (Tallinn Manual 2.0).

Article 36 Additional Protocol I to the Geneva Conventions (API) facilitates the application of IHL to contemporary developments by demanding the compliance of 'new' means and methods of warfare with established IHL principles. However, as is the case with all cyber operations, the applicability of IHL rules to deepfakes proves to be no clear-cut a matter. The most comprehensive, yet non-binding, international guideline on cyber warfare is the Tallinn Manual 2.0, which unfortunately only sparsely touches on the implication of different forms of disinformation in armed conflict and makes no mentioning of deepfakes. According to Rule 80 Tallinn Manual 2.0, the *existence* of an armed conflict is a prerequisite for the applicability of IHL to cyber operations. Thus, deepfakes which are employed during an ongoing armed conflict, are governed by the same IHL rules as the 'traditional' means and methods of warfare employed in that conflict – notwithstanding that these rules might not be sufficient or appropriate in the context of information warfare.

Perfidy and ruses of war

The deception of an adversary in armed conflicts by dissemination of false information is a contemporary method of warfare. In principle, deepfakes are nothing but a more sophisticated, hyper-realistic continuation of this practice, as the following examples show:

Example 1: State A produces a deepfake of a representative of the International Committee of the Red Cross inviting both adversaries to e. g. peace talks. State A sends the deepfake to the military commander of State B with the intent to attack State B's representatives on their way to the faked meeting.

Example 2: State A produces a deepfake in which State B's military commander orders the armed forces of State B to retreat from strategically important cities. State A then spreads the deepfake video to the armed forces of State B via social media, with the intent of gaining a military advantage.

IHL offers black letter law to determine which acts of deception are permitted. According to Article 37(1) API, perfidious acts – those which invite particular confidence in the adversary and intend to betray that confidence – are prohibited. Example 1 clearly falls within the scope of Article 37(1) API and is therefore prohibited.

In contrast, under Article 37(2) API "acts which are intended to mislead an adversary or to induce him to act recklessly but which infringe no rule of international law applicable in armed conflict and which are not perfidious because they do not invite the confidence of an adversary" constitute permitted ruses of war. Article 37(2) API names "misinformation" as an example of a permissible ruse. Rule 123 Tallinn Manual 2.0 cites *inter alia* the spreading of disinformation causing an adversary to erroneously believe a false appearance of what is actually happening, and "bogus orders purporting to have been issued by the enemy commander" as prime examples of permissible ruses. Accordingly, example 2 will most likely be considered as permissible ruse, provided that no other IHL is violated.

Disinformation, civilians, and the notion of "attack"

In the assessment of the legality of deepfakes, their impact on the civilian population is pivotal.

The duty of constant care (Article 57(1) API), the principles of distinction (Article 48 API) and proportionality (Article 51(5)(b) API), and the prohibition of acts whose primary aim is to spread terror among the civilian population (Article 51(2) API, Article 13(2) Additional Protocol II to the Geneva Conventions), can possibly limit the lawful usage of deepfakes in armed conflicts.

Example 3: State A produces a deepfake depicting a conversation between State B's president and military commander about an imminent nuclear attack on the capital city of State C. State A disseminates the deepfake through social media platforms to primarily cause panic across the civilian populations of States B and C.

While the duty of constant care and the principle of distinction apply in all military operations, the principle of proportionality and the prohibition of acts whose primary aim is to spread terror, only govern "attacks", acts of violence, or threats thereof. This is where specific difficulties regarding the applicability to and classification of deepfakes arise. There is virtually no case law to define what constitutes an 'attack' in the context of cyber conflicts. "Attack", as per Article 49(1) API, "means acts of

violence against the adversary […].” Rule 92 Tallinn Manual 2.0 states that violence “must be considered in the sense of violent consequences and is not limited to violent acts.” However, according to Rule 98, a Twitter message “sent out in order to cause panic, falsely indicating that a highly contagious and deadly disease is spreading rapidly throughout the population […] is neither an attack […] nor a threat thereof […]” and consequently does not violate Article 51(2) API. But can fake news in no circumstance be an attack, act of violence, or threat thereof?

While the exemplary tweet disseminates ‘news’ without claiming state involvement, the deepfake illustrated in example 3 is of an entirely different quality. As a deepfake can only be recognized as such with great difficulty and therefore will – at least initially – be perceived as authentic, the effects on State C and its civilian population will certainly not be any less grave than those of a real announcement of a nuclear attack. However, Rule 92 Tallinn Manual 2.0 deems operations causing “inconvenience or irritation” without foreseeably resulting in injury of individuals or damage of physical objects lawful. Thus, under the current framework, due to a lack of foreseeable injury, example 3 would be considered a lawful non-attack, although the resulting panic and terror could be tremendous.

Similarly, the distribution of the deepfake under example 2 via social media raises questions regarding its conformity with the principle of distinction and the prohibition of indiscriminate attacks (Article 51(4) API). The interconnectedness of cyberspace makes it practically impossible to strictly distinguish between civil and military uses of (social) media and reasonably foreseeable that any deepfake can (accidentally) fall into the hands of civilians. Regardless of whether the objectives of the deepfake are civilian or military, the notion of ‘attack’ is decisive. Rule 93 Tallinn Manual 2.0 (Distinction) stipulates that operations directed at civilians are only prohibited when they amount to an ‘attack’; operations directed at military objectives must comply with the principle of proportionality – which only applies to attacks. According to Rule 105 Tallinn Manual 2.0, cyber weapons creating a chain of events beyond the control of the attacker are indiscriminate by nature. While social media spreads information uncontrollably, as long as the deepfake does not *foreseeably* cause injury or damage – which the deepfake of example 2 clearly does not – it is not indiscriminate. Accordingly, also example 2 would be lawful.

Conclusion

Even though IHL can somewhat grasp the concept of disinformation in warfare due to the longstanding practice, the existing legal framework is not equipped to appropriately react to the dimensions deepfakes add to the equation. Due to technological advances, it has become necessary for the law to differentiate between the available forms and contents of ‘fake news’. The most pressing issues, namely the subsumption of deepfakes under the notions of ‘attack’, ‘act of violence’, or threat thereof, and the requirements of foreseeability and degree of possible harm, are in dire need for clarification, as both are determinate for the applicability of pertinent IHL norms and the protection of civilians. First steps in countering deepfakes could be the installment of safeguard-mechanisms, e.g. digital watermarks, and including the concept of deepfakes in future international manuals to facilitate the discourse between states and ascertain current *opinio iuris*.